Otto-Friedrich-Universität Bamberg

Professur für Angewandte
Informatik insb. Kognitive Systeme

Masterarbeit

im Studiengang Angewandte Informatik
der Fakultät Wirtschaftsinformatik
und Angewandte Informatik
der Otto-Friedrich-Universität Bamberg

Semantisches Matching von Freizeitaktivitäten mittels Wikipediabasierter Kategorisierung

Semantic matching of leisure activities by Wikipedia based categorisation

Autor:

Matthias Düsel

Betreuer:

Prof. Dr. Ute Schmid

Bamberg, Dezember 2013

Abstract

This paper examines how Wikipedia can be used to compute semantic similarity. The goal is to create a matching system, in which users enter a leisure activity and get back entries with similar activities as a result. In this work the focus is on the meaning of the entries. Therefore, matches will be based on semantic similarity and not just on same words in two entries. In a purely lexical comparison, there is no match between „*I would like to go to the opera*“ and “*I visit Don Giovanni today*“. These entries are semantically related, since Don Giovanni is an opera.

After introducing chapters about semantic similarity this thesis describes how the semantic of an item is determined by using Wikipedia and Wiktionary. Therefore, tags, which are the results of Wikipedia requests, are added to each entry. In this context the category system of Wikipedia is being used. Tags are keywords which describe an object, in this case an entry. The system, which was developed as a part of this thesis, uses titles of Wikipedia articles or Wikipedia categories as tags. Text mining of user entries is managed by using Wiktionary.

Furthermore, various approaches to find matches between entries are described. Overall, in the software seven different methods are implemented. They range from a simple comparison of tags to the search for common top-level categories in the Wikipedia category graph. Finally, strengths and weaknesses of using Wikipedia are analyzed and possible approaches for improvement are discussed.

Zusammenfassung

Diese Arbeit untersucht, inwiefern Wikipedia genutzt werden kann, um semantische Ähnlichkeit zu berechnen. Das Ziel ist es, ein Matching-System zu erstellen, in welches ein Benutzer eine Freizeitaktivität eingibt und daraufhin Einträge mit ähnlichen Aktivitäten als Ergebnis zurück bekommt. Hierbei liegt das Augenmerk auf der Bedeutung der Einträge. Die Matches sollen also aufgrund ihrer semantischen Ähnlichkeit entstehen und nicht nur, weil zwei Einträge das selbe Wort enthalten. Bei einem rein lexikalischem Vergleich gibt es kein Match zwischen „*Ich würde gerne in die Oper gehen*“ und „*Ich sehe mir heute Don Giovanni an*“. Semantisch sind diese Einträge jedoch verwandt, da Don Giovanni eine Oper ist.

Nach einführenden Kapiteln über semantische Ähnlichkeit wird in der vorliegenden Masterarbeit beschrieben, wie mithilfe von Wikipedia und Wiktionary die Semantik eines Eintrags ermittelt wird. Dazu werden zu jedem Eintrag Tags hinzugefügt, welche die Ergebnisse von Wikipedia-Anfragen sind. Dabei wird das Kategorie-System von Wikipedia verwendet. Tags sind Schlagwörter, welche ein Objekt, hier einen Eintrag, beschreiben. Im Falle des hier entwickelten Systems sind Tags Titel von Wikipedia-Artikeln oder Wikipedia-Kategorien. Das Textmining der Benutzereinträge wird mittels Wiktionary bewältigt.

Anschließend werden verschiedene Ansätze beschrieben, um Matches zwischen Einträgen zu finden. Insgesamt wurden im Programm, welches im Laufe dieser Arbeit erstellt wurde, sieben verschiedene Methoden implementiert. Diese reichen von einem einfachen Vergleich der Tags bis hin zur Suche nach gemeinsamen Oberkategorien im Wikipedia-Kategoriegraph. Abschließend werden Stärken und Schwächen der Verwendung von Wikipedia analysiert und mögliche Ansätze zur Verbesserung diskutiert.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

Vielen Studien zufolge leiden ältere Menschen häufig an Einsamkeit. Dies kann viele Ursachen haben, wie beispielsweise der Umzug in ein Altenheim oder der Verlust der eigenen Mobilität. Für diese Leute ist es schwer, Menschen mit den selben Interessen und Hobbies zu finden. Das Ziel dieser Arbeit ist es, ein System zu entwickeln, welches Menschen zusammenführt, die die selben Freizeitaktivitäten ausführen. So können sich zwei Personen, welche beide gerne Konzerte besuchen, für einen gemeinsamen Konzertbesuch verabreden. Außerdem können Menschen, welche eine Aktivität nicht mehr alleine ausführen können, Hilfe finden. Kann jemand beispielsweise nicht mehr eigenständig Auto fahren und möchte dennoch eine Oper in einer anderen Stadt besuchen, gibt es die Möglichkeit sich einer anderen Person anzuschließen. Ebenfalls denkbar wären Matches zwischen Einträgen von Benutzern, die Hilfe suchen, und solchen, die Hilfe anbieten. Ein Match existiert immer dann, wenn das System einem Eintrag eines Benutzers den Eintrag eines anderen Benutzers zuordnet.

Im Gegensatz zu rein lexikalischen Matching-Systemen, welche gleiche Wörter in zwei Einträgen finden, soll hier ein System entwickelt werden, das auf der Semantik, also der Bedeutung, der Einträge basiert. So sollen sich beispielsweise zwei Personen finden, von welchen eine gerne in die Oper geht und die andere gerne ein Konzert besuchen würde. Besitzen die beiden Einträge keine übereinstimmenden Wörter, wird trotzdem ein Match gefunden, da „*Oper*“ und „*Konzert*“ über die gemeinsame Oberkategorie „*Musik*“ semantisch verwandt sind.

In Kapitel 2 werden die theoretischen Grundlagen für die Bestimmung von semantischer Ähnlichkeit beschrieben. Dazu werden verschiedene Ansätze vorgestellt. Zuerst wird der Schlüsselwortvergleich als rein lexikalische Vergleich behandelt. Anschließend wird auf verschiedene Ontologie- und Taxonomiebasierte Ansätze eingegangen und herausgestellt, warum diese in semantischer Hinsicht besser geeignet sind. Im letzten Abschnitt dieses Kapitels werden Wikipediabasierte Ansätze zur Bestimmung von semantischer Ähnlichkeit beschrieben. Hier wird auch auf das Kategoriensystem von Wikipedia eingegangen, welches im weiteren Verlauf verwendet wird.

Das dritte Kapitel behandelt die in dieser Arbeit umgesetzten Ansätze aus Kapitel 2. Die Bestimmung von semantischer Ähnlichkeit zwischen zwei Einträgen basiert auf Tags. Zunächst wird beschrieben, wie diese Tags erstellt werden. Danach werden die verschiedenen Arten des Matchings vorgestellt. Für das im Rahmen dieser Masterarbeit erstellte Java-Programm wurden sieben verschiedene Ansätze zur Bestimmung von semantischer Ähnlichkeit implementiert. Das letzte Unterkapitel beschreibt, wie vom Programm aus auf Wikipedia und Wiktionary zugegriffen wird.

Kapitel 4 liefert detaillierte Ausführungen zur Implementierung. Dabei wird sowohl auf die Komplexität als auch auf Vor- und Nachteile der verschiedenen Matching-Methoden eingegangen. Außerdem werden die Matches anhand eines Beispiels mit elf Einträgen analysiert und bewertet. Dazu werden jeweils Recall und Precision berechnet. Zusätzlich werden allgemeine Stärken und Schwächen des entwickelten Programms erklärt.

Im letzten Kapitel werden die erreichten Ziele beschrieben und die aufgetretenen Probleme zusammengefasst. Außerdem werden Ansatzpunkte für weitere Verbesserungen gegeben.

2 Semantische Ähnlichkeit

Semantische Ähnlichkeit von zwei Texten beschreibt wie nahe sich der Inhalt, also die Bedeutung, beider Dokumente sind. Im Folgenden werden verschiedene Ansätze beschrieben, wie die semantische Ähnlichkeit ermittelt werden kann.

2.1 Lexikalischer Ansatz

Der im nächsten Kapitel vorgestellte Schlüsselwortvergleich ist ein rein lexikalisches Verfahren, welches auf keine semantische Hintergrundinformation zugreift. Dadurch werden nur genaue Übereinstimmungen gefunden. Wie später noch beschrieben wird, sind rein lexikalische Verfahren dadurch nur bedingt für die Bewertung von semantischer Ähnlichkeit geeignet.

2.1.1 Schlüsselwortvergleich

Um die semantische Ähnlichkeit zweier Dokumente zu ermitteln kann der Schlüsselwortvergleich eingesetzt werden. Dabei wird überprüft, ob einzelne Wörter in beiden Texten vorkommen. Dazu werden zunächst aus beiden Dokumenten Schlüsselwörter extrahiert, welche den jeweiligen Text gut repräsentieren. Anschließend werden diese verglichen. Durch die Anzahl der gemeinsamen Schlüsselwörter wird auf den Grad der semantischen Ähnlichkeit der Dokumente geschlossen.

Hierbei wird nicht auf die eigentliche Semantik, also die Bedeutung des Textes, eingegangen. Es werden lediglich die Schlüsselwörter Buchstabe für Buchstabe verglichen. So können beispielsweise zwei Dokumente von digitalen Speichermedien handeln, wobei im einen Text ausschließlich von CDs und im anderen von DVDs gesprochen wird. Da die Schlüsselwörter „CD“ und „DVD“ lexikalisch nicht gleich sind, wird die hohe semantische Verwandtschaft nicht erkannt. [1]

Zur automatischen Erkennung von Schlüsselwörtern gibt es verschiedene Ansätze. Im Allgemeinen wird zunächst eine Stoppwortelimination durchgeführt. Dabei werden nicht relevante Wörter wie Artikel oder Konjunktionen gelöscht. Das geschieht über den lexikalischen Vergleich mit einer Liste, die solche Stoppwörter enthält. Als zweiter Schritt kann eine Stammformreduktion durchgeführt werden. Hierdurch werden alle Wörter auf ihren Wortstamm zurückgeführt. So werden lexikalische Unterschiede, welche durch Einzahl und Mehrzahl oder durch Deklinationen und Konjugationen entstehen, ausgeglichen. Die ist jedoch in manchen Sprachen sehr schwierig und funktioniert beispielsweise im Deutschen aufgrund der vielen Unregelmäßigkeiten nicht so effektiv wie im Englischen.

Nach diesen Vorbereitungen werden die Schlüsselwörter extrahiert. Es wird davon ausgegangen, dass ein Wort für einen Text umso relevanter ist, je häufiger es darin enthalten ist [3]. Besitzt man mehrere Dokumente, also einen Textkorpus, so ist ein Schlüsselwort sehr viel diskriminierender für einen Text, wenn es in diesem sehr häufig und in den anderen Texten selten vorkommt. Die dadurch erhaltene Gewichtung wird als Term-Frequency Inverted-Document-Frequency (TF-IDF) bezeichnet. [10]

2.1.2 Wortmehrdeutigkeiten

In fast allen Sprachen treten Mehrdeutigkeiten von Wörtern, welche auch als Ambiguitäten, Homonyme oder Polyseme bezeichnet werden, auf. Beispielsweise kann „plant“ im Englischen sowohl „Pflanze“ als auch „Fabrik“ bedeuten [4]. Ein klassisches Beispiel im Deutschen ist „die Bank“ für eine Sitzgelegenheit oder ein Geldinstitut. Polysemie grenzt sich von der Homonymie durch einen gemeinsamen etymologischen Ursprung der Wortbedeutung ab [6]. Diese Unterscheidung ist jedoch für die Ermittlung der semantischen Nähe unbedeutend.

Vergleicht man zwei Dokumente auf ihre semantische Ähnlichkeit durch den lexikalischen Schlüsselwortvergleich, kann es dadurch zu verfälschten Ergebnissen kommen. Handelt ein Text von einer Parkbank und ein anderer von einem Geldinstitut und beides Mal wird nur das Wort „Bank“ verwendet, so ergibt der Schlüsselwortvergleich fälschlicherweise eine hohe semantische Ähnlichkeit. Wie solche Wortmehrdeutigkeiten durch semantische Hintergrundinformation aufgelöst werden, zeigen die folgenden Kapitel.

2.2 Ontologie- und Taxonomiebasierte Ansätze

Um die semantische Ähnlichkeit von lexikalisch unterschiedlichen Wörtern zu berechnen, ist es nötig auf die Bedeutung, also die Semantik, der Wörter einzugehen. Dazu ist zusätzliches Hintergrundwissen nötig. Eine Möglichkeit, die Nähe der Bedeutung von zwei Wörtern zu ermitteln, ist die Verwendung von Taxonomien oder Ontologien. Eine Taxonomie ist eine Baumstruktur, die Begriffe zueinander in Beziehung stellt. Hierbei liegt eine hierarchische Ordnung vor. Es gibt also eine eindeutige Wurzel, welche das Hyperonym für alle anderen Begriffe ist. Das bedeutet, dass alle Kindknoten Unterbegriffe, sogenannte Hyponymien, vom Begriff in der Taxonomie-Wurzel sind. Der Elternknoten ist dementsprechend immer der Oberbegriff des aktuell betrachteten Knotens. In einer Taxonomie existieren nur is-a-Beziehungen. Das bedeutet, ein Kindknoten gehört zur Kategorie des Elternknotens. Tiefer in der Struktur gelegene Knoten sind demzufolge Unterkategorien der höher gelegenen Knoten. Ein Beispiel wäre „Hund“ ist ein „Tier“. [12]

Ontologien sind komplexer als Taxonomien. Diese müssen nicht zwingend hierarchisch geordnet sein. Die Verbindungen der einzelnen Knoten sind nicht vorgeschrieben. So könnten zum Beispiel die Begriffe „Mohnblume“ und „Blut“ über die Beziehung „ist rot“ verbunden sein. Um Verbindungen zwischen den Knoten zu erstellen wird auf die Eigenschaften der einzelnen Begriffe eingegangen.

2.2.1 Pfadbasierte Ähnlichkeit

Durch die Repräsentation der Begriffe in einer Baumstruktur lässt sich die semantische Nähe von zwei Wörtern ermitteln. Dazu wird der Abstand im Graph, also die Pfadlänge, berechnet. In einer Taxonomie bedeutet dies, je kürzer die Pfadlänge zwischen zwei Knoten ist, desto ähnlicher ist sich die Semantik der beiden Begriffe. Ist der Pfad sehr kurz, so liegen die beiden Knoten im selben Teilbaum des Graphs und die gemeinsame Wurzel, welche der Kategorie entspricht, der beide Begriffe angehören, ist nicht weit entfernt. Ebenfalls wäre es denkbar, das ein Begriff eine Unterkategorie des anderen ist.

Rada et al.[7] entwickelten eine Formel zur Berechnung der semantischen Ähnlichkeit basierend auf der Pfadlänge. Die semantische Distanz $sim(c_1, c_2)$ zwischen zwei Begriffen entspricht der minimalen Kantenanzahl zwischen den beiden Knoten in der Taxonomie. Begriffe in einer Taxonomie werden auch als Konzepte bezeichnet, daher werden die Abkürzungen c_1 und c_2 verwendet.

$$sim(c_1, c_2) = plength(c_1, c_2)$$

Wendet man die Berechnung auf komplette Dokumente, beziehungsweise ihrer Menge von Schlüsselwörtern an, so wird die Pfadlänge über alle Kombinationen der Begriffe beider Texte gemittelt. Dazu wird folgende Formel verwendet:

$$sim(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{i=1}^{|C_1|} \sum_{j=1}^{|C_2|} sim(c_i, c_j) \qquad c_i \in C_1, c_j \in C_2$$

C_1 und C_2 sind hierbei die Menge aller Begriffe der beiden zu vergleichenden Dokumente.

Die Ergebnisse der somit ermittelten Ähnlichkeitswerte hängen sehr stark von der Qualität der verwendeten Taxonomie ab.

Wu und Palmer [13] stellen eine normierte Berechnung der semantischen Ähnlichkeit, welche auf der Pfadlänge in der Baumstruktur basiert, vor.

$$sim_{WP}(c_1, c_2) = \frac{2 * depth(lcs(c_1, c_2))}{plength(c_1, lcs(c_1, c_2)) + plength(c_2, lcs(c_1, c_2)) + 2 * depth(lcs(c_1, c_2))}$$

In diesem Ansatz wird zunächst die Tiefe des sogenannten lowest common subsumer (lcs) ermittelt. Dieser Knoten subsumiert die Eigenschaften beider Konzepte c_1 und c_2. Der lcs ist also das nächstgelegene Hyperonym beider Begriffe. Anschließend wird der Abstand zwischen diesem Knoten und den beiden Konzepten berechnet und mit dem lcs normiert.

In Abbildung 1 ist eine vereinfachte Taxonomie dargestellt. Es soll die semantische Ähnlichkeit $sim_{WP}(c_1, c_2)$ zwischen dem Knoten 4 und dem Knoten 7 berechnet werden. Der lowest common subsumer ist Knoten 3. Dieser besitzt eine Tiefe von 2, welche dem Abstand zur Wurzel (Knoten 1) entspricht. Der Abstand von Knoten 4 (c_1) zu Knoten 3 (lcs) beträgt 1 und der zu Knoten 7 (c_2) ist 2. Somit ergibt sich für die semantische Distanz ein Wert von 0,5714.

$$\begin{aligned} sim_{WP}(K4, K7) &= \frac{2 * N3}{N1 + N2 + 2 * N3} \\ &= \frac{2 * depth(K3)}{plength(K4, K3) + plength(K7, K3) + 2 * depth(K3)} \\ &= \frac{2 * 2}{1 + 2 + 2 * 2} \\ &= 0,5714 \end{aligned}$$

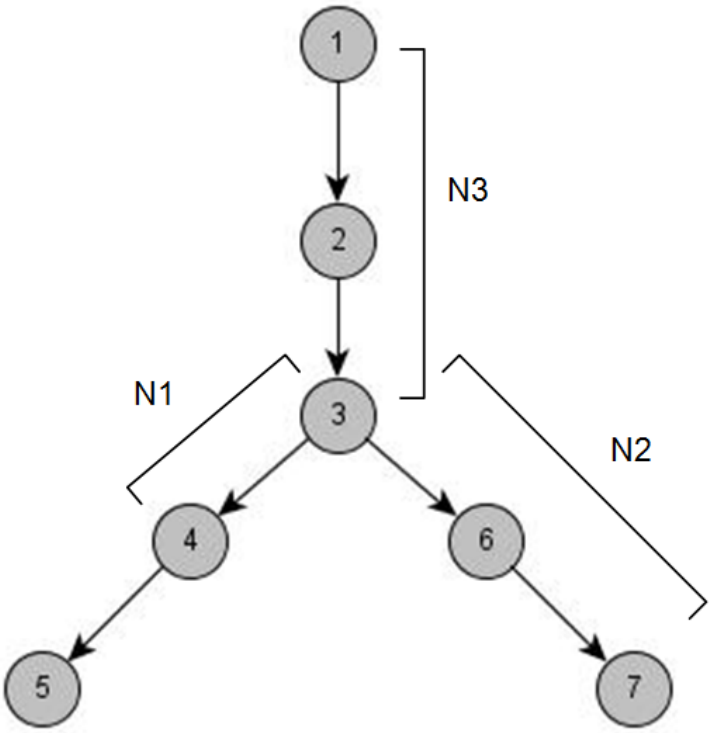

Abbildung 1: Vereinfachte Taxonomie

Der Vorteil dieser Berechnung gegenüber der zuvor beschriebenen semantischen Distanz von Rada et al. ist, dass hier normierte Werte zwischen 0 und 1 vorliegen. Außerdem wird die komplette Struktur der Taxonomie miteinbezogen. So könnten beispielsweise durch einen zusätzlichen Faktor tiefer in der Struktur gelegene semantische Ähnlichkeiten höher gewichtet werden. Damit würden Ähnlichkeiten zwischen speziellen Konzepten, welche automatisch tiefer in der Struktur liegen, stärker bewertet werden als allgemeine, welche oben in der Baumstruktur angesiedelt sind.

2.2.2 Informationsgehaltbasierte Ähnlichkeit

Informationsgehaltbasierte Ansätze gehen davon aus, dass sich zwei Konzepte umso ähnlicher sind, je mehr Information sie gemeinsam haben. Der Informationsgehalt ist eine logarithmische Kennzahl, welche Auskunft über die Häufigkeit beziehungsweise die Wahrscheinlichkeit, dass ein Konzept in einem Text vorkommt, gibt.

Nach Ross [9] lässt sich der Informationsgehalt als negativer Logarithmus darstellen. Dies ist sinnvoll, da so der Informationsgehalt steigt, wenn die Wahrscheinlichkeit, dass ein Konzept vorkommt, sinkt. Der Informationsgehalt ist also umso höher, je spezieller das Konzept ist. Die Wahrscheinlichkeit für ein Konzept c wird dabei wie folgt berechnet.

$$p(c) = \frac{count(c)}{n}$$

Der Ausdruck $count(c)$ bezeichnet die Anzahl der Vorkommen des Konzepts c im gesamten

Textkorpus. Hier ist also eine Menge von Dokumenten nötig, um diese Wahrscheinlichkeit zu berechnen. Dazu kann beispielsweise ein Standardkorpus verwendet werden. Resnik [8] bezieht sich beispielsweise auf den Brown Corpus of American English. Er zählt hierbei das Vorkommen eines Nomens im gesamten Korpus. Die Wahrscheinlichkeit dafür, dass es ein Konzept gibt, welches dieses Nomen enthält, ist umso großer, je häufiger das Nomen vorkommt. Wird dabei der Zähler $count(c)$ für ein Konzept erhöht, so wird gleichzeitig der Zähler für alle übergeordneten Konzepte erhöht. Es gilt also $p(c_1) \leq p(c_2)$, wenn c_1 tiefer in einer is-a-Taxonomie liegt als c_2. $count(c)$ ist demnach die Summe aller Vorkommen von c und aller untergeordneten Konzepte. n ist die Anzahl aller gezählten Instanzen im gesamten Korpus.

Resnik [8] entwickelte draus die Formel für die semantische Ähnlichkeit sim_R.

$$sim_R(c_1, c_2) = -\log p(lcs(c_1, c_2))$$

Setzt man hier die Formel für die Wahrscheinlichkeit $p(c)$ ein, ergibt sich folgende Gesamtformel:

$$sim_R(c_1, c_2) = -\log \frac{count(lcs(c_1, c_2))}{n}$$

Da der Wurzelknoten der Taxonomie zu jedem Konzept ein übergeordnetes Konzept ist, wird dessen Zähler immer ebenfalls erhöht. Demnach beträgt $count(Wurzel)$ n und $p(Wurzel)$ ist 1. Sind zwei Konzepte nur über den Wurzelknoten miteinander verbunden, so ist das niedrigste gemeinsame übergeordnete Konzept die Wurzel. Daraus ergibt sich eine semantische Ähnlichkeit von $-\log \frac{n}{n} = -\log 1 = 0$.

Lin [2] erweiterte den Ansatz von Resnik. Sein Ähnlichkeitsmaß stützt sich auf drei Hypothesen.

1. Die Ähnlichkeit von A und B ist abhängig von ihren Gemeinsamkeiten. Je mehr Gemeinsamkeiten sie besitzen, desto ähnlicher sind sich die beiden Objekte.
2. Die Ähnlichkeit von A und B ist abhängig von ihren Unterschieden. Je mehr Unterschiede sie besitzen, desto weniger ähnlich sind sich die beiden Objekte.
3. Die größte Ähnlichkeit zwischen A und B herrscht, wenn A und B identisch sind. Dies gilt unabhängig von ihren Gemeinsamkeiten.

Die Gemeinsamkeiten von A und B misst Lin mit dem Informationsgehalt. Wie bereits im vorherigen Abschnitt beschrieben, ist der Informationsgehalt der negative Logarithmus der Wahrscheinlichkeit, dass eine Information enthalten ist.

$$I(common(A, B)) = -\log p(common(A, B))$$

common(A,B) steht hierbei für die Gemeinsamkeiten von A und B. In einer is-a-Taxonomie könnten dies Beispielsweise die gemeinsamen übergeordneten Konzepte sein.

Die Unterschiedlichkeit von A und B, bezogen auf die zweite Hypothese, beschreibt er mit der Differenz aus dem Informationsgehalt der vollständigen Beschreibung beider Konzepte $description(A, B)$ und dem Informationsgehalt ihrer Gemeinsamkeiten.

$$I(description(A, B)) - I(common(A, B))$$

Das hieraus gefolgerte Ähnlichkeitsmaß setzt den Informationsgehalt der Aussage, die die Gemeinsamkeiten von A und B beschreibt und den Informationsgehalt der Aussagen, die A und B vollständig beschreiben, ins Verhältnis.

$$sim_L(A, B) = \frac{\log p(common(A, B))}{\log p(description(A, B))}$$

Für Taxonomien lässt sich daraus folgende Formel ableiten:

$$sim_L(c_1, c_2) = \frac{2 * \log p(lcs(c_1, c_2))}{\log p(c_1) + \log p(c_2)}$$

Die Gemeinsamkeiten beider Konzepte werden durch das in der Baumstruktur tiefste gemeinsame übergeordnete Konzept (lcs) dargestellt. Als Beschreibung der Objekte dient jeweils das Konzept c.

2.3 Wikipediabasierte Ähnlichkeit

Ein sich immer weiter verbreitender Ansatz semantische Ähnlichkeit zu bestimmen, ist die Verwendung von Wikipedia. Die freie Enzyklopädie ist in vielerlei Hinsicht interessant. Durch die Vielzahl an Autoren wächst sie täglich weiter und ist dadurch nicht begrenzt, wie beispielsweise künstlich erzeugte Taxonomien. Außerdem ist in Wikipedia durch die Struktur automatisch semantische Information enthalten. Autoren können einem Artikel beispielsweise eine oder mehrere Kategorien zuordnen und so das durch den Artikel beschriebene Wort in ein semantisches Netz, den sogenannten Kategoriebaum, einordnen. Zusätzlich sind die Artikel untereinander durch Links verknüpft. Dies gibt ebenfalls darüber Auskunft, wie stark unterschiedliche Artikel semantisch miteinander verbunden sind. Das so entstehende Netzwerk wird als Artikelgraph bezeichnet. Ein weiteres Merkmal, das die semantischen Eigenschaften der Wikipedia zeigt, sind die Begriffsklärungsseiten. Hier werden Mehrdeutigkeiten von Wörtern, wie sie in 2.1.2 beschrieben werden, aufgelöst. Außerdem gibt es noch die sogenannten Weiterleitungsseiten, welche Synonyme auflösen. Diese leiten den Benutzer beispielsweise automatisch von „Auto“ zu „Automobil“ oder von „Geige“ zu „Violine“ weiter. All diese semantische Information kann genutzt werden, um die semantische Nähe zweier Begriffe zu ermitteln.

2.3.1 Linkbasierte Ähnlichkeit

Eine Methode zur Bestimmung der semantischen Ähnlichkeit, welche nur auf den Links innerhalb der Wikipedia-Artikel basiert, stellt Milne in seinem Artikel vor [5]. Für zwei zu

vergleichende Terme werden zunächst die passenden Artikel aus Wikipedia gesucht. Falls ein Term einen eigenen Artikel hat, also ein direktes Matching vorliegt, wird dieser Artikel verwendet. Falls nicht, wird über Weiterleitungen und Begriffsklärungsseiten der passende Artikel bestimmt. Anschließend werden diese Artikel in ein Vektormodel überführt. Dazu werden nur die Links innerhalb der Artikel betrachtet. Ähnlich dem aus dem Information Retrieval bekannten Verktorraummodel, bei welchem die einzelnen Terme mit dem zuvor beschriebenen TF-IDF-Verfahren gewichtet werden, erhalten die Links unterschiedliche Gewichtungen. Dabei wird die Auftrittswahrscheinlichkeit für diesen Link in der kompletten Wikipedia berücksichtigt. Ist t die gesamte Anzahl an Wikipedia-Artikeln, so ist w der gewichtete Wert für einen Link $a \to b$ zwischen den Artikeln a und b.

$$w(a \to b) = |a \to b| * \log\left(\sum_{x=1}^{t} \frac{t}{x \to b}\right)$$

Der gewichtete Wert für einen Link ist also das Produkt aus der Anzahl des Links im aktuellen Artikel und der inversen Wahrscheinlichkeit für das Auftreten irgendeines Links zum Zieldokument. Für zwei Artikel x und y ergeben sich nach der Berechnung aller n auftretenden Links $l_i|i = 1..n$ folgende Vektoren:

$$x = (w(x \to l_1), w(x \to l_2), ..., w(x \to l_n))$$
$$y = (w(y \to l_1), w(y \to l_2), ..., w(y \to l_n))$$

Als Ähnlichkeitsmaß wird der Winkel zwischen diesen Vektoren berechnet. Dabei besitzen Artikel, welche eine übereinstimmende Liste von Links haben, einen Winkel von 0°. Besteht keinerlei Überlappung der Links beider Artikel und herrscht damit laut Milne keine semantische Ähnlichkeit, so beträgt der Winkel 90°. Werden für einen Term zu Beginn aufgrund von Wortmehrdeutigkeiten mehrere Artikel gefunden, so wird der Vektorvergleich paarweise mit allen Artikeln und dem Artikel des zweiten Begriffs durchgeführt. Als Endergebnis wird nur der kleinste Wert, also der Artikel, der dem anderem am ähnlichsten ist, ausgewählt. So werden Ambiguitäten, welche nicht mit der gesuchten Wortbedeutung übereinstimmen, automatisch verworfen.

2.3.2 Kategoriensystem

Der sogenannte Kategoriebaum von Wikipedia ist im eigentlichen Sinne kein Baum, da Mehrfachzuweisungen von Kategorien erlaubt sind und der Graph somit multihierarchisch sein kann. Trotzdem ist es das Ziel, die Kategorien hierarchisch vom Allgemeinen zum Besonderen zu ordnen. Laut [14] ist dieser Graph geeignet, um darauf graphbasierte Ähnlichkeitsbestimmungen, wie sie in Abschnitt 2.2 beschrieben sind, anzuwenden. Strube und Ponzetto [11] kommen zu dem Schluss, dass eine beschränkte Tiefensuche mit der Suchtiefe 4 zu den besten Ergebnissen führt, den least common subsumer (lcs) zu finden und so den kürzesten Pfad zwischen zwei Kategorien zu ermitteln. In Abbildung 2 ist zu erkennen, dass der Kategoriegraph mit dem Artikelgraph verbunden ist. Hier ist auch verdeutlicht, dass einem Wikipedia-Artikel eine oder mehrere Kategorien zugeordnet sein können. Das Ähnlichkeitsmaß wird zu allen Kategoriepaaren beider Artikel berechnet und anschließend das beste Ergebnis gewählt.

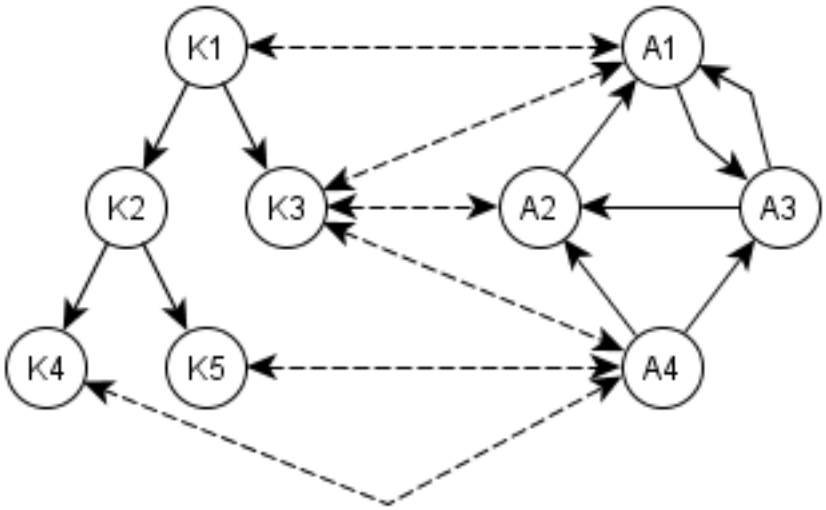

Abbildung 2: Verknüpfung von Kategorie- und Artikelgraph

Sollen beispielsweise die Artikel A2 und A4 verglichen werden, wird zunächst festgestellt, dass A2 nur die Kategorie K3 zugeordnet ist. A4 hingegen besitzt die Kategoriemenge [K3, K4, K5]. Es werden also die Ähnlichkeitsmaße für (K3, K3), (K3, K4) und (K3, K5) berechnet. Da die Kategorie K3 übereinstimmt, wird dieses als bestes Ergebnis ausgewählt. Die Artikel besitzen eine hohe semantische Ähnlichkeit. [14]

3 Matching durch semantische Ähnlichkeit

Die Verwendung von Wikipedia ist durch ihren enormen Umfang gerechtfertigt. Seit Mai 2001 sind 1.634.158 Artikel in deutscher Sprache entstanden[1]. Zwar ist die Qualität eines von Experten künstlich erstellten Thesaurus, wie beispielsweise GermaNet, sicherlich besser, allerdings sind diese in Umfang und Aktualität begrenzt. Bei einem auf kollaborativem Wissen basierenden System wie Wikipedia wird die Qualität durch die gegenseitige Überprüfung der Autoren untereinander, also der Community, gewährleistet. Dem gegenüber steht GermaNet mit 111.361 lexikalische Einheiten[2]. Dies ist zwar ein klar strukturiertes Wissensnetz, welches von Linguisten erstellt wurde, allerdings ist dadurch die Entstehung teuer und damit verbunden die Größe begrenzt. Weiterhin können die oft großen Abstände der Aktualisierungszyklen solcher linguistischen Datenbanken nicht mit der Dynamik eines kollaborativen Systems mithalten. So fehlen beispielsweise Einträge von aktuellen Ereignissen oder Eigennamen.

Um zwei Einträge auf der semantischen Ebene zu vergleichen muss zuerst die Semantik jedes Eintrags bestimmt werden. In dieser Arbeit werden dazu sogenannte Tags verwendet. Tags sind Schlagwörter, welche ein Objekt, hier einen Eintrag, beschreiben. Diese Tags sind also Stichwörter, welche die Semantik eines Eintrags bezeichnen.

3.1 Erstellung der Tags

Um geeignete Tags für eine Aktivitätsbeschreibung zu erhalten, muss der Eintrag in das Matchingsystem in mehrfacher Hinsicht bearbeitet werden. Zunächst stellt sich die Frage, welche Wortklassen für aussagekräftige Tags verwendet werden sollten. Hier fällt die erste Wahl ganz klar auf Substantive. Diese besitzen die größte Aussagekraft über den Inhalt eines Textes. Zusätzlich werden hier noch Verben berücksichtigt. Allerdings müssen zuvor noch Wörter, welche allgemein sehr häufig verwendet werden, herausgefiltert werden. Sie sagen wenig über die Semantik aus und dienen nicht zur Diskriminierung gegenüber anderen Texten. Dies geschieht durch eine so genannte Stoppwortliste.

Nachdem der Eintrag des Matchingsystems von Interpunktionen und anderen Sonderzeichen bereinigt wurde, werden nicht relevante Wörter herausgefiltert. Dazu wird, wie in Abschnitt 2.1.1 beschrieben, ein lexikalischer Vergleich mit einer Stoppwortliste durchgeführt. Im System, das im Rahmen dieser Masterarbeit entwickelt wurde, wird hierzu die Liste der Université de Neuchâtel[3] verwendet. Stimmen Wörter des Eintrags mit Wörtern der Stoppwortliste überein, werden diese gelöscht. Aufgrund der Thematik des Matchers wurden die Wörter „brauchen“, „suchen“, „helfen“ und „begleiten“ zur Stoppwortliste hinzugefügt. Diese kommen erwartungsgemäß sehr häufig vor und geben keine Auskunft über die Semantik eines Eintrags.

[1] http://de.wikipedia.org/, Stand 26.09.2013
[2] http://www.sfs.uni-tuebingen.de/lsd/, Stand April 2013
[3] http://members.unine.ch/jacques.savoy/clef/index.html, Stand 20.09.2013

3.1.1 Wiktionary

Da in Wikipedia die Grundform eines Wortes als Sucheingabe verwendet werden muss, wird eine Art Textmining als Vorverarbeitung durchgeführt. Enthält ein zu kategorisierender Eintrag beispielsweise das Wort „Töchter“, so liefert eine Anfrage bei Wikipedia keinen direkten Treffer. Denkbar wäre es, einen Stemmer einzusetzen, welcher jedes Wort auf seinen Wortstamm reduziert. So führt zum Beispiel der Snowball-Stemmer[4] die Wörter „Tochter“ und „Töchter“ auf den gemeinsamen Wortstamm „tocht“ zurück. Allerdings besitzt Wikipedia keinen Eintrag mit diesem Titel. Es ist also ein Verfahren notwendig, welches Plural- oder Diminutivformen sowie Deklinationen und Konjunktionen erkennt und diese in die Grundform des Wortes überführt. Ein solches Verfahren wird Lemmatisierung genannt. Aus den gleichen Gründen, aus denen in dieser Arbeit Wikipedia verwendet wird, wird hier zur Lemmatisierung Wiktionary verwendet.

Wiktionary ist ein freies Online-Wörterbuch mit 340.211 Einträgen[5]. Es ist, wie Wikipedia, ein Projekt der Organisation Wikimedia Foundation, Inc., das durch freie Autoren kollaborativ erweitert werden kann. Durch die Verwendung von Wiktionary in dieser Arbeit, gibt es keine Lizenzprobleme, wie es beispielsweise bei kommerziellen Natural Language Processing (NLP) Systemen, welche oft einen Lemmatisierer beinhalten, vorkommen kann. Außerdem ist Wiktionary dynamisch und wächst durch die Kollaboration kontinuierlich weiter.


Töchter

Töchter (Deutsch) [Bearbeiten]

Deklinierte Form [Bearbeiten]

Worttrennung:
Töch·ter

Aussprache:
IPA: ['tœçtɐ]
Hörbeispiele: Töchter (Info)

Grammatische Merkmale:
- Nominativ Plural des Substantivs Tochter
- Genitiv Plural des Substantivs Tochter
- Akkusativ Plural des Substantivs Tochter

Töchter ist eine flektierte Form von Tochter.
Alle weiteren Informationen zu diesem Wort findest du im Eintrag Tochter.
Bitte nimm Ergänzungen deshalb auch nur dort vor.

Abbildung 3: Lemmatisierung mit Wiktionary

Wird in Wiktionary nach dem Wort „*Töchter*“ gesucht, so wird erkannt, dass dieses eine flektierte Form des Wortes „*Tochter*“ ist. Zu erkennen ist dies, wie in Abbildung 3 veranschaulicht, im grün gefärbten Kasten am unteren Rand der Wiktionary-Seite.

Anschließend an die Stoppworteliminationon wird mit jedem verbleibenden Wort des Eintrags eine Suchanfrage an Wiktionary gestellt. Das Ziel hiervon ist es zum Einen, die Wortklasse zu ermitteln. Adverbien und Adjektive werden aus der verbleibenden Wortmenge gelöscht. Substantive gehen später mit voller Gewichtung in die Tag-Menge ein. Verben werden nur mit halbem Gewicht berücksichtigt, da sie weniger Aussagekraft über den Inhalt eines Textes besitzen. Zum anderen wird, wie in Abbildung 3 dargestellt, das jeweilige Wort in seine Grundform überführt. Das Wort „*Töchter*“ wird also durch „*Tochter*“ ersetzt. Da ein Wort, welches nicht in seiner Grundform war, nicht erkannt wurde, muss anschließend erneut geprüft werden, ob sich das Wort in der Stoppwortliste befindet oder ob es sich um ein Adjektiv oder ein Adverb handelt.

Enthält Wiktionary keinen Eintrag zu einem Wort, so handelt es sich dabei mit hoher

[4] http://snowball.tartarus.org/
[5] http://de.wiktionary.org/, Stand 26.09.2013

Wahrscheinlichkeit um einen Eigennamen. Um auch Eigennamen zu erkennen, welche aus zwei Wörtern zusammen gesetzt sind, werden Kombinationen aus den umliegenden Wörtern ebenfalls gespeichert. Lautet der Eintrag in das Matchingsystem beispielsweise „*Ich suche eine Begleitung für ein Spiel der Brose Baskets*“, ergibt sich nach der Stoppwortelimination die Wortmenge [*Begleitung, Spiel, Brose, Baskets*]. In Wiktionary gibt es keinen Treffer für das Wort „*Brose*“. Aus diesem Grund werden zusätzlich die Strings „*Spiel Brose*“ und „*Brose Baskets*“ gespeichert. So wird die Wahrscheinlichkeit erhöht, dass bei der folgenden Wikipedia-Anfrage zweielementige Eigennamen, welche beispielsweise auch aus Vor- und Nachname bestehen können, gefunden werden. Diese zweielementigen Tags werden zuerst als Suchbegriff in Wikipedia verwendet. Liegt hier ein Treffer vor, so werden alle Tags, welche eines der beiden Wörter enthalten, aus der Menge gelöscht. In dem hier genannten Beispiel enthält Wikipedia keinen Artikel für „*Spiel Brose*“, aber einen für „*Brose Baskets*“. Daher lautet die neue Tag-Menge [*Begleitung, Spiel, Brose Baskets*]. Die Tags „*Spiel Brose*“, „*Brose*“ und „*Baskets*“ werden gelöscht.

Um dieses Verfahren zu vereinfachen und sicher zu gehen, dass keine falschen Tags gelöscht werden, ist es möglich, dass der Endbenutzer Eigennamen selbst kennzeichnet. Dies geschieht mittels Anführungszeichen. Lautet der Eintrag eines Benutzers „*Ich suche eine Begleitung für ein Spiel der „Brose Baskets““*, so werden die einzelnen Tags „*Brose*“ und „*Baskets*“ von Anfang an nicht aufgenommen. Dadurch können auch mehrelementige Eigennamen oder Namen, welche einem Stoppwort entsprechen würden, berücksichtigt werden.

3.1.2 Wikipedia

Zusätzlich zu den schon vorhandenen Tags sollen mit Hilfe von Wikipedia weitere Tags hinzugefügt werden, welche ein späteres Matching mit anderen Einträgen erleichtern sollen. Dabei wird sowohl auf das Kategoriensystem als auch auf die Weiterleitungsfunktion von Wikipedia zurückgegriffen. Außerdem wird Wikipedia genutzt, um den Tags verschiedene Gewichtungen zu geben.

Lautet ein Eintrag in das Matchingsystem beispielsweise „*Ich sehe mir am Sonntag Don Giovanni an. Wer begleitet mich?*“, ergibt sich aus der in den letzten Abschnitten beschriebenen Verarbeitung die Wortmenge

[*Giovanni, Don Giovanni, Giovanni begleitet, Don, Sonntag, begleiten*].

Zu jedem dieser Wörter wird nun eine Wikipedia-Anfrage gestartet. Existiert ein Artikel zu einem Begriff, so werden aus den angegebenen Kategorien neue Tags generiert. Dabei wird auf das in Abschnitt 2.3.2 beschriebene Kategoriensystem zurückgegriffen. In Abbildung 4 sind die Kategorien eines Wikipedia-Artikels am Ende der Seite zu sehen. Der gefundene Artikel zu „*Don Giovanni*“ besitzt mehrere Kategorieeinträge.

Das Verfahren für zweielementige Tags wurde am Ende des letzten Abschnitts bereits beschrieben. Durch einen Treffer bei der zweielementigen Suchanfrage „*Don Giovanni*“ werden die Elemente „*Giovanni*“, „*Giovanni begleitet*“ und „*Don*“ gelöscht und es wird ein Tag „*Don Giovanni*“ mit der vollen Gewichtung, also 1.0, erstellt. Sind keine Kategorien vorhanden, erhält der Begriff nur ein Gewicht von 0.5. In den meisten dieser Fälle existiert

zu diesem Suchbegriff entweder überhaupt kein Wikipedia-Artikel oder es handelt sich um eine Begriffsklärungsseite.

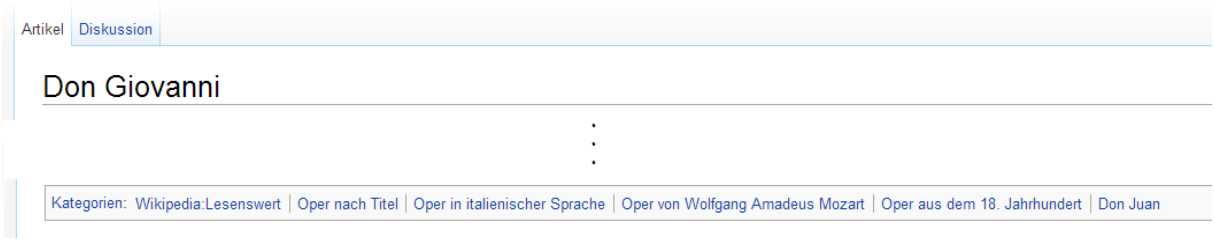

Abbildung 4: Kategorien eines Wikipedia-Artikels

In Wikipedia muss zwischen Artikeln und Kategorien unterschieden werden. Häufig sind die Titel von Artikeln auch Kategorien, aber nicht immer. Der Artikel „Oper“ besitzt beispielsweise ebenfalls die Kategorie „Oper“. In Abbildung 5 ist der Unterschied dargestellt.

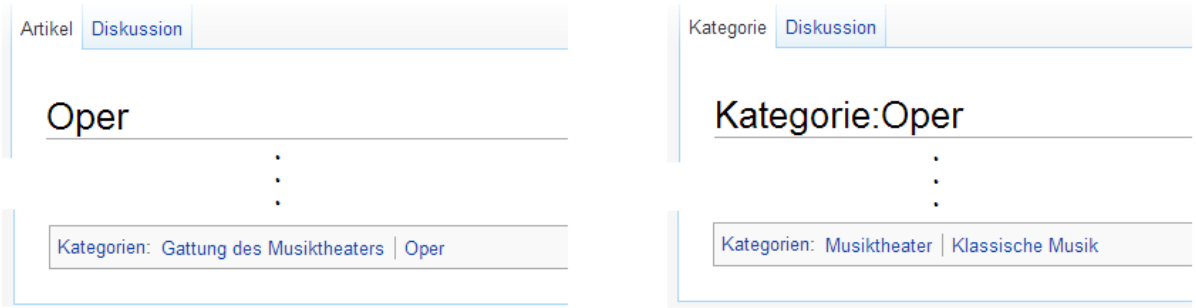

Abbildung 5: Unterschied zwischen Artikeln und Kategorien

Wie zu erkennen ist, können die Kategorien von Artikel- und Kategorieseiten voneinander abweichen. Die Kategorien von Kategorieseiten entsprechen immer den Oberkategorien, also den Elternknoten im Kategoriegraph. Bei einem Artikel hingegen kann der Autor beliebige Kategorien angeben. Zur Gewinnung von neuen Tags, werden daher sowohl Artikel- als auch Kategorie-Seiten berücksichtigt. Dadurch werden sowohl Kategorien als auch Wörter aus dem Eintrag, welche einen Wikipedia-Artikel besitzen, aber keiner Kategorie entsprechen, als Tags aufgenommen.

Im nächsten Schritt wird versucht, Kategorien zusammenzufassen. Dazu werden zunächst die Stoppwörter aus den Kategorien gelöscht. Eigene Kategorien von Wikipedia wie beispielsweise „*Wikipedia:Lesenswert*“ und alle Hauptkategorien[6] werden ebenfalls entfernt. Anschließend werden je zwei Kategorien verglichen und der kleinste gemeinsame Teilstring gefunden. Stimmen ein oder mehrere zusammenhängende Wörter überein, so wird dies als neue Kategorie aufgenommen und die beiden ursprünglichen Kategorien gelöscht. Dies geschieht allerdings erst, sobald alle Kategorien untereinander verglichen worden sind, so dass alle Übereinstimmungen der verschiedenen Kategorien gefunden werden. Danach startet der Vergleich von vorne, so dass alle Kategorien, welche einen gemeinsamen Teil besitzen, zu diesem zusammengefasst werden.

[6] http://de.wikipedia.org/wiki/Kategorie:!Hauptkategorie

So werden im Beispiel aus Abbildung 4 alle Kategorien, welche das Wort „*Oper*“ enthalten zu einem Tag zusammengefasst. Eine solche zusammengefasste Kategorie erhält das volle Gewicht. Einzelne Kategorien werden als Tag mit halbem Gewicht aufgenommen.

Besitzen zwei Kategorien mehr als einen übereinstimmenden Teilstring, werden alle diese Teilstrings erkannt und als Tags aufgenommen. Durch dieses Verfahren werden wichtige Inhalte der Kategorien hervorgehoben, da sie durch die Zusammenfassung das doppelte Gewicht erhalten. Zusätzlich verringert sich dadurch die Anzahl der Tags, was sich positiv auf die Komplexität des späteren Matchings auswirkt.

Wird an Wikipedia eine Anfrage mit einem Verb gestellt, erfolgt häufig eine Weiterleitung zum Artikel mit dem dazugehörigen Substantiv. In diesem Beispiel führt der Begriff „*begleiten*“ zur Seite „*Begleitung*“. Zu solchen weitergeleiteten Treffern werden Tags mit halbem Gewicht erzeugt. Es ergibt sich abschließend die Tag-Menge

[*Oper, 1.0* | *Don Juan, 0.5* | *Don Giovanni, 1.0* | *Begleitung, 0.5* | *Sonntag, 1.0*]

Dadurch dass bei einer Weiterleitung immer das Ziel als Tag übernommen wird, steigt die Wahrscheinlichkeit, dass semantisch verwandte Einträge den selben Tag enthalten, zu welchem sie von unterschiedlichen Ausgangswörtern weitergeleitet wurden.

3.2 Matching durch den Vergleich der Tags

Wird ein neuer Eintrag in das Matchingsystem eingegeben, so wird anschließend automatisch eine Suche nach zu diesem Eintrag passenden, schon vorhandenen Einträgen durchgeführt. Eine solche Suche wird als Matching bezeichnet und kann zusätzlich jeder Zeit manuell ausgelöst werden. Hierzu wurden verschiedene Ansätze implementiert, welche im Folgenden beschrieben und in Kapitel 4 evaluiert werden.

3.2.1 Gleiche Tags

Im Standardfall werden die Tags der einzelnen Einträge mit den Tags des ausgewählten Eintrags verglichen. Enthält ein Eintrag den gleichen Tag, so wird das Tag-Gewicht des aktuell betrachteten Eintrags mit dem Tag-Gewicht des gefunden Eintrags multipliziert. Das Gewicht wird für alle übereinstimmenden Tags aufsummiert. Das Match mit der höchsten Gewichtung besitzt die meisten übereinstimmenden relevanten Tags und ist somit semantisch am ähnlichsten. Deutlich wird das am folgendem Beispiel.

ID	Eintrag	Autor
0	Ich sehe mir am Sonntag Don Giovanni an. Wer begleitet mich?	Person2
1	Ich brauche jemanden der mir beim Einkaufen hilft.	Person1
2	Wer möchte mit mir in ein Konzert gehen?	Person4
3	Ich gehe gerne in die Oper.	Person5
4	Ich gehe heute Abend in die Oper und suche noch jemanden der mitkommt.	Person3

Abbildung 6: Einträge des Matchingsystems

Die in Abbildung 6 zu sehenden Einträge mit den IDs 0 bis 3 sind bereits vorhanden. Der letzte Eintrag „*Ich gehe heute Abend in die Oper und suche noch jemanden der mitkommt.*“ wurde gerade hinzugefügt. Hierfür wurden folgende Tags mit ihren Gewichtungen erstellt:

[*Abend, 1.0* | *Oper, 1.0* | *Tageszeit, 0.5* | *Gattung Musiktheaters, 0.5* | *Klassische Musik, 0.5* | *Musiktheater, 0.5*] .

Das Matching ergibt für den Eintrag mit der ID 3 das höchste Gesamtgewicht. Die Tags dieses Eintrags lauten:

[*Oper, 1.0* | *Gattung Musiktheaters, 0.5* | *Klassische Musik, 0.5* | *Musiktheater, 0.5*] .

Das Gesamtgewicht g_G beträgt

$$g_G = \sum_{i=1}^{n} g_{1,i} \times g_{2,i} = 1.0 \times 1.0 + 0.5 \times 0.5 + 0.5 \times 0.5 + 0.5 \times 0.5 = 1.75$$

wobei $g_{1,i}$ das Gewicht des i-ten Tags des ersten Eintrags und $g_{2,i}$ das Gewicht des i-ten Tags des zweiten Eintrags ist. n entspricht der Anzahl an gleichen Tags der beiden Einträge.

Der Eintrag mit der ID 0 ist ebenfalls ein Match, da gleiche Tags vorhanden sind. Dieser enthält die Tags:

[*Oper, 1.0* | *Don Giovanni, 1.0* | *Sonntag, 1.0* | *Don Juan, 0.5* | *Sehen, 0.5* | *Visuelle Wahrnehmung, 0.5* | *Auge, 0.5*| *Wahrnehmung, 0.5*| *Wochentag, 0.5*].

Es stimmt nur der Tag „*Oper*“ überein.

Das Ergebnis des Matchings wird als Tabelle mit absteigendem Gesamtgewicht geliefert.

ID	Eintrag	Autor	Gewicht
3	Ich gehe gerne in die Oper.	Person3	1.75
0	Ich sehe mir am Sonntag Don Giovanni an. Wer begleitet mich?	Person0	1.0

Abbildung 7: Ergebnis des Matchings

3.2.2 Teilweise Übereinstimmung von Tags

Zusätzlich zu der genauen Übereinstimmung von Tags ist es auch möglich, nach teilweisen Übereinstimmung zu suchen. Bei diesem Vergleich müssen die zwei Tags nicht komplett gleich sein, sondern es genügt ein gemeinsamer Teilstring. Dabei muss der Teilstring mindestens 5 Buchstaben besitzen und am Anfang oder am Ende eines Wortes stehen. Dadurch wächst die Chance, Teile zusammengesetzter Wörter zu ermitteln. Würde man dies nicht berücksichtigen, so gäbe es viele Übereinstimmungen, die nicht gewollt sind. Ein Beispiel wäre die Übereinstimmung von „*Anstalt*“ mit „*Ver**anstalt**ung*“. Die Mindestlänge von 5 ist dadurch begründet, dass viele Wörter der deutschen Sprache gleiche Endungen haben. So würde beispielsweise die häufige Endung „*-tung*“ zu vielen falsch positiv erkannten Matchings führen. Das selbe gilt für Präfixe am Wortanfang.

Wird in das Beispiel aus Abbildung 14 der Eintrag „*Ich interessiere mich für alle Arten von Musikveranstaltungen.*“ eingefügt, so ergibt das Standardmatching keinen Treffer. Durch das hier beschriebene Verfahren mit Teilstrings wird allerdings das Wort „*Musik*“ in den Tags der Einträge mit den IDs 3 und 4 erkannt. Der neue Eintrag enthält den Tag „*Musikveranstaltung, 0.5*“. Die beiden gefundenen Einträge enthalten jeweils die Tags „*Gattung Musiktheaters, 0.5*“, „*Musiktheater, 0.5*“ und „*Klassische Musik, 0.5*“, welche über die Wikipedia-Suche zu dem Wort „*Oper*“ zustande kommt. Die Gewichte werden jeweils multipliziert. Da es sich hier um kein direktes Match handelt, wird das Gewicht halbiert. In diesem Beispiel ergibt sich demnach jeweils ein Gesamtgewicht von 0.75.

ID	Eintrag	Autor	Gewicht
3	Ich gehe gerne in die Oper.	Person3	0.75
4	Ich gehe heute Abend in die Oper und suche noch jemanden der mitkommt.	Person4	0.75

Abbildung 8: Ergebnis des Matching mit Teilstrings

Das Teilstring-Matching kann über das Menü des Matchers ein- und ausgeschaltet werden, wie in Abbildung 9 zu sehen ist. Die damit zusammenhängende Erhöhung der Laufzeit und der falsch positiven Matchings wird in Kapitel 4 untersucht.

Abbildung 9: Optionales Teilstring-Matching

3.2.3 Kategorisierung von Tags

Wird zu einem Eintrag kein Match gefunden, besteht die Möglichkeit, dessen Tags, erneut zu kategorisieren und anschließen ein erneutes Matching durchzuführen. Dieses Verfahren wird im Folgenden auch als *intensive Suche* bezeichnet. Die Tags eines Eintrags bestehen aus Substantiven und Verben des Eintrags sowie den dazugehörigen Ergebnissen der Wikipedia-Kategorisierung, wie sie in Abschnitt 3.1.2 beschrieben wird. Da schon vorhandene Tags nicht erneut hinzugefügt werden, findet keine erneute Kategorisierung der Substantive und Verben statt. Bei jedem anderem Tag handelt es sich um einen Titel eines Wikipedia-Artikels, eine Kategorie aus Wikipedia oder einen Tag, welcher mehrere Kategorien zusammenfasst (vgl. „*Oper*“ in Abschnitt 3.1.2).

Die Kategorisierung dieser Tags erfolgt analog zur ersten Verarbeitung. Ähnliche Kategorien werden ebenfalls wieder zusammengefasst. Verfolgt man weiterhin die Beispiele aus den vorherigen Kapiteln mit den beiden Einträgen

„Ich sehe mir am Sonntag Don Giovanni an. Wer begleitet mich?“

[*Oper, 1.0* | *Don Giovanni, 1.0* | *Sonntag, 1.0* | *Don Juan, 0.5* | *Sehen, 0.5* | *Visuelle Wahrnehmung, 0.5* | *Auge, 0.5*| *Wahrnehmung, 0.5*| *Wochentag, 0.5*]

und

„Ich interessiere mich für alle Arten von Musikveranstaltungen.“
[*Art, 1.0* | *Musikveranstaltungen, 0.5* | *Interesse, 0.5*]

, so ergibt weder das Standardmatching noch der Vergleich von Teilstrings eine Übereinstimmung. Wird allerdings der Tag „*Oper*“, welcher in diesem Fall eine zusammengefasste Kategorie ist, erneut kategorisiert, kommt der Tag „*Gattung Musiktheaters, 0.5*“ hinzu. Dadurch ist ein Übereinstimmung anhand von Teilstrings möglich und es wird das gemeinsame Wort „Musik“ erkannt. Somit wird die semantische Verwandtschaft zwischen den beiden Einträgen aufgedeckt.

Da die Kategorisierung über Wikipedia je nach Internetverbindung viel Zeit in Anspruch nimmt, gehört diese Methode nicht zum Standardmatching, sondern wird nur angewandt, falls keine Treffer gefunden werden. Aus dem selben Grund wird hier nur für den Eintrag, zu dem passende Partner gefunden werden sollen, ein erneutes Matching der Tags durchgeführt. Dadurch entstehen zwar Zeitersparnisse, allerdings werden aber auch nicht alle Matches gefunden. Wird beispielsweise eine solche *intensive Suche* für den zweiten Eintrag durchgeführt, so wird kein neuer Tag hinzugefügt, welcher ein Matching mit dem ersten Tag ermöglicht. Die Wikipedia-Anfragen zu den Wörtern „*Musikveranstaltungen*“, „*Art*“ und „*Interessieren*“ ergeben keine neuen Tags, die mit denen des ersten Eintrags übereinstimmen.

Eine Lösung hierfür wäre es, zu jedem im Matchingsystem vorkommenden Eintrag die Tags erneut zu kategorisieren. Ob das geschehen soll, kann im Matchingsystem manuell ausgewählt werden (Abbildung 10). Dies nimmt allerdings je nach Anzahl der Einträge und der jeweiligen Tags viel Zeit in Anspruch. Außerdem können dadurch falsch positive Matchings entstehen, da sich die Tags semantisch zu sehr vom ursprünglichen Eintrag entfernen. Dieser Aspekt wird in der Evaluation im folgenden Kapitel genauer betrachtet.

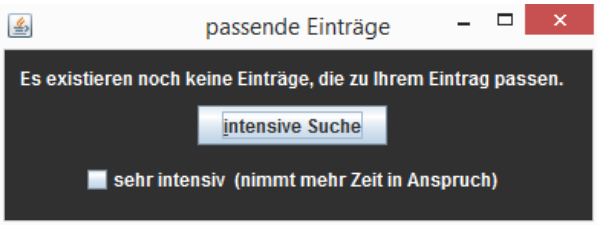

Abbildung 10: Intensive Suche mit der Möglichkeit, die Tags aller Einträge erneut zu kategorisieren

3.2.4 Tiefensuche im Wikipedia-Kategoriegraph

Eine zusätzliche Erweiterung ist die Suche nach einem least common subsumer (lcs), also der nächst gelegenen gemeinsamen Kategorie. Dieses Verfahren wird in Abschnitt 2.3.2 beschrieben. Werden zwei Einträge miteinander verglichen, werden deren Tags so lange

neu kategorisiert, bis eine Übereinstimmung vorliegt. Anschließend wird der kleinste Wert, also der kürzeste Pfad durch den Kategoriegraph als Ähnlichkeitsmaß der beiden Einträge bestimmt. Dabei werden nur komplett übereinstimmende Tags berücksichtigt und keine Teilstrings.

Wie Strube und Ponzetto in [11] beschreiben, ist es sinnvoll, die Suche nach dem lcs auf eine Tiefe von 4 zu beschränken, da die Kategorien, welche weit oben im Graph liegen, zu eng verknüpft sind. Dabei handelt es sich um allgemeine Oberkategorien. Im Gegensatz zur *intensiven Suche*, bei welcher analog zur erstmaligen Tag-Erstellung Artikel- und Kategorie-Seiten berücksichtigt werden, wird hier nur der Kategoriegraph verwendet. Es werden also in jedem Schritt die direkten Oberkategorien der aktuellen Kategorien hinzugefügt. Die Suche nach dem kürzestem Pfad findet also in einer Taxonomie statt.

Durch dieses Verfahren werden sehr viele Tags neu kategorisiert und der Aufwand steigt exponentiell, weil für jeden Tag immer wieder mehrere neue Tags hinzugefügt werden. Dadurch wird viel Zeit für die Wikipedia-Anfragen benötigt. Ob dies durch das Ergebnis gerechtfertigt ist, wird in Kapitel 4 evaluiert.

3.3 Wikimedia-Zugriff über Java

In dem hier entwickelten System wird zur Laufzeit auf die beiden Wikimedia-Systeme Wikipedia und Wiktionary zugegriffen. Dazu wird die offene Java-Bibliothek *gwtwiki Java Wikipedia API (Bliki engine)*[7] verwendet. Es wird absichtlich kein XML-Dump, also ein komplettes Abbild zu einem bestimmten Zeitpunkt, benutzt. Dies hätte zwar zur Laufzeit Vorteile, was die Zugriffszeit betrifft, allerdings gibt es auch Nachteile. Zum einen wird mehr Speicherplatz benötigt. Die Größe eines XML-Dumps der deutschen Wikipedia beträgt ohne Bilder ca. 3 GB. Ein weiterer Aspekt ist die Aktualität. Da sich Wikimedia-Projekte dynamisch täglich weiterentwickeln, wäre das System größtenteils nicht auf dem aktuellen Stand und es müssten häufige Updates durchgeführt werden. Neue XML-Dumps werden ca. alle zwei bis vier Wochen bereitgestellt.[8] Da die Zielgruppe ältere Menschen sind, welche möglicherweise den Umgang mit Computern nicht sehr gewöhnt sind, wird hier ein System ohne Updates bevorzugt.

Um die Einträge aller Benutzer zusammenzubringen, ist ohnehin eine Internetverbindung notwendig. Daher entsteht durch die Verwendung des Echtzeitzugriffs auf die Wikimedia-Seiten kein Nachteil. Je nach Geschwindigkeit der Internetverbindung, kann es hier zu Verzögerungen kommen. Vor allem bei Verwendung der *intensiven Suche*(Abschnitt 3.2.3) oder der Suche nach dem lcs (Abschnitt 3.2.4), kann es zu einer hohen Anzahl von Anfragen kommen. Da die Anfragen größtenteils sehr gezielt und im Normalfall von geringer Anzahl sind, überwiegen die Vorteile. Eine genaue Analyse wird in Kapitel 4 durchgeführt.

[7] https://code.google.com/p/gwtwiki/
[8] http://de.wikipedia.org/wiki/Hilfe:Download

4 Umsetzung und Evaluation

In diesem Kapitel werden verschiedenen Methoden des Matchings, welche im letzten Kapitel vorgestellt wurden, genauer untersucht. Dabei wird zuerst auf die Umsetzung und Komplexität eingegangen. Anschließend werden Recall und Precision der gefundenen Ergebnisse berechnet. Im Folgenden wird der Eintrag, zu welchem passende Matchings gefunden werden sollen, als Ausgangseintrag bezeichnet. Die unterschiedlichen Matching-Ansätze können im Programm für jeden Eintrag per Kontextmenü ausgeführt werden.

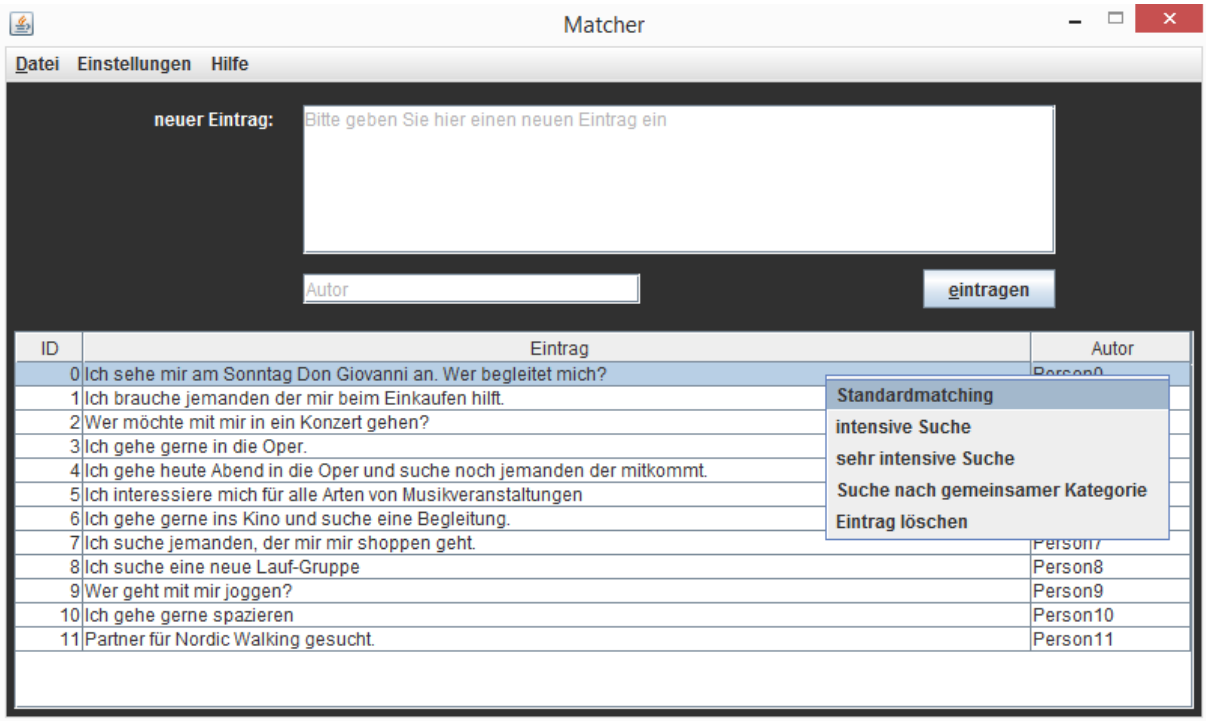

Abbildung 11: Matching per Kontextmenü

4.1 Gleiche Tags

Beim im vorherigen Kapitel als Standardmatching bezeichneten Verfahren werden alle Tags aller vorhandenen Einträge durchsucht. Es handelt sich also um eine dreifach geschachtelte For-Schleife in der für jeden Eintrag jeder Tag des Ausgangseintrag mit jedem Tag des Eintrags verglichen wird. Die äußere For-Schleife durchläuft die Einträge, die zweite die Tags des Ausgangseintrags und die innerste die Tags des aktuell betrachteten Eintrags. Dadurch ergibt sich folgende Komplexität:

$$O(n) = (n-1) \times t \times \frac{\sum_{i=1}^{n-1} k}{n-1} = t \times \sum_{i=1}^{n-1} k$$

Dabei ist n die Anzahl der Einträge. Da der Ausgangseintrag nicht mit sich selbst verglichen wird, wird hier mit $(n-1)$ gerechnet. t ist die Anzahl der Tags des Ausgangseintrags. Bei k handelt es sich um die Anzahl der Tags des jeweiligen Eintrags, mit dem der Ausgangseintrag gerade verglichen wird.

4.2 Teilweise Übereinstimmung von Tags

Ob überprüft werden soll, ob Tags teilweise übereinstimmen, wird durch einen booleschen Wert gesteuert, welcher vom Benutzer geändert werden kann (vgl. Abbildung 9). Dieser wird durch eine einfache If-Abfrage überprüft. Die Hauptkomplexität ist demnach die selbe wie bei der Überprüfung auf vollständig identische Tags. Zusätzlich wird für jedes Durchlaufen der innersten For-Schleife, also für jeden Vergleich von zwei Tags eine Funktion aufgerufen, welche den kleinsten gemeinsamen Teilstring ermittelt. Dabei werden die Tags Buchstabe für Buchstabe miteinander verglichen. Da die Tags im Allgemeinen sehr kurz sind, entstehen hier keine übermäßigen Kosten.

4.3 Kategorisierung von Tags

Bei dieser Art des Matchings, welches bereits als *intensive Suche*(3.2.3) beschrieben wurde, werden vor dem Vergleich der Tags, die bereits vorhandenen Tags erneut über Wikipedia kategorisiert. Dadurch wird die Anzahl der Tags und damit die Chance auf eine Übereinstimmung erhöht. Hierbei kann der Benutzer wählen, ob nur die Tags des Ausgangseintrags erneut kategorisiert werden sollen oder die Tags aller Einträge. Je nach Gesamtanzahl der vorhandenen Einträge kann dies zu einer beträchtlichen Komplexität- und Laufzeiterhöhung führen.

Bei der erneuten Kategorisierung werden, wie bei der ursprünglichen Erstellung der Tags, Artikel- und Kategorie-Seiten von Wikipedia verwendet. Dies bedeutet, dass zu jedem Tag zwei Wikipedia-Anfragen gestellt werden, mit welchen die neuen Kategorien ermittelt werden. Zusätzlich wird im Voraus noch eine Anfrage gestellt, um den Inhalt der Wikipedia-Seite abzufragen und so Begriffsklärungs- und Weiterleitungs-Seiten zu erkennen. Anfragen nach Kategorien und Inhalt sind in der hier verwendeten API zwei unterschiedliche Funktionen. Handelt es sich um eine Begriffsklärungs- oder Weiterleitungs-Seite, so wird für diesen Tag keine Artikel-Anfrage mehr gestellt. Eine Kategorie-Anfrage kann dennoch ein Ergebnis erzielen, wie in Abbildung 12 zu erkennen ist.

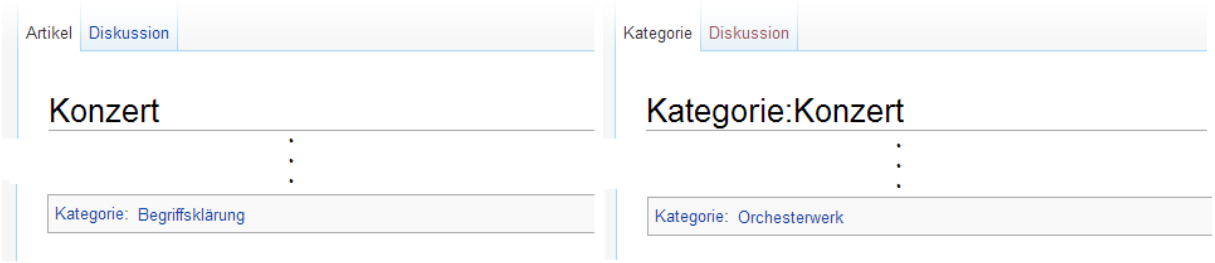

Abbildung 12: Unterschied zwischen Artikel- und Kategorie-Anfrage

Wurden die Tags des Ausgangseintrags beziehungsweise die Tags aller Einträge kategorisiert, wird das Standardmatching durchgeführt. Dabei wird je nach Benutzereinstellung mit oder ohne Teilstring-Übereinstimmung gearbeitet.

4.4 Tiefensuche im Wikipedia-Kategoriegraph

Bei der Tiefensuche im Wikipedia-Kategoriegraph wird ein Matching zwischen zwei Einträgen durch das Hinzufügen von Oberkategorien der Tags erreicht. Dadurch wird die kürzeste Verbindung im Kategoriegraph gefunden. Dazu werden in jedem Schritt die zu diesem Zeitpunkt aktuellen Tags des Ausgangseintrags und die des aktuell betrachteten Eintrags als Kategorie-Anfrage an Wikipedia gestellt. Die Artikel-Seiten werden hierbei nicht betrachtet, da die Kategorien von Artikeln nicht hierarchisch aufgebaut sind. Dieser Unterschied wurde in Abbildung 5 dargestellt.

Wie bereits in Abschnitt 3.2.4 beschrieben, wird diese Prozedur durchgeführt, bis ein Match gefunden wurde oder nach viermaligem Durchlaufen noch kein Ergebnis vorliegt. Bei einer größeren Suchtiefe, sind die Oberkategorien zu weit von einander entfernt und es handelt sich bei einem Match sehr wahrscheinlich um zu allgemeine Kategorien.

Um die Anfragen an Wikipedia, und damit die Laufzeit dieses Matchings, zu minimieren werden in jedem Schritt nur die im letzten Schritt neu hinzugefügten Kategorien kategorisiert. Dadurch wird ebenfalls vermieden, dass die hier verwendete Wikipedia-API (vlg. 3.3) nicht alle Ergebnisse zurückliefert. Dazu kommt es, wenn zu viele Anfragen gleichzeitig gestellt werden.

Enthält ein Eintrag den Tag „*Oper*“ und ein anderer den Tag „*Konzert*“, so wird ein Matching über die gemeinsame Oberkategorie „*Musik*“ gefunden. Abbildung 13, welche mit Hilfe des CatGraph-Werkzeuges vom Wikimedia-Mitglied Dapete[9] erstellt wurde, zeigt die Oberkategorien beider Tags.

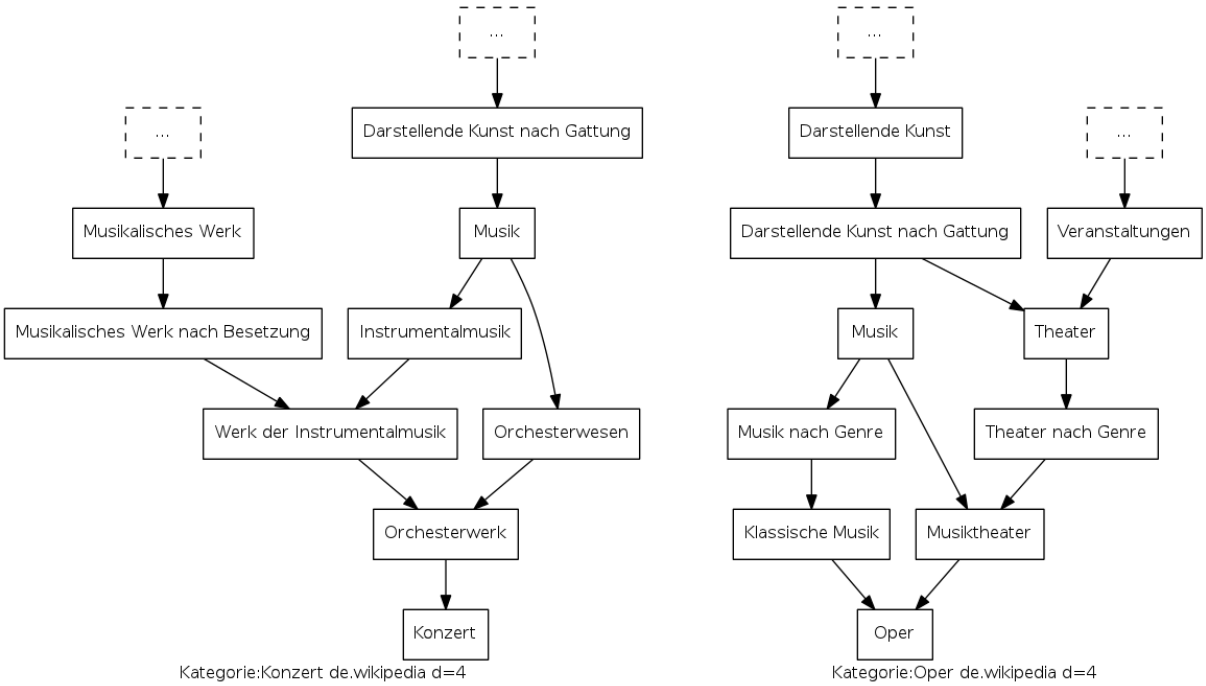

Abbildung 13: Gemeinsame Oberkategorie

[9]http://toolserver.org/ dapete/catgraph/, 30.11.2013

4.5 Vergleich der Matching-Ansätze

Die beschriebenen Matching-Ansätze sollen nun anhand von Recall und Precision beschrieben werden. Dazu werden die elf Einträge aus Abbildung 14 verwendet.

ID	Eintrag	Autor
0	Ich sehe mir am Sonntag Don Giovanni an. Wer begleitet mich?	Person0
1	Ich brauche jemanden der mir beim Einkaufen hilft.	Person1
2	Wer möchte mit mir in ein Konzert gehen?	Person2
3	Ich gehe gerne in die Oper.	Person3
4	Ich gehe heute Abend in die Oper und suche noch jemanden der mitkommt.	Person4
5	Ich interessiere mich für alle Arten von Musikveranstaltungen	Person5
6	Ich gehe gerne ins Kino und suche eine Begleitung.	Person6
7	Ich suche jemanden, der mir mir shoppen geht.	Person7
8	Ich suche eine neue Lauf-Gruppe	Person8
9	Wer geht mit mir joggen?	Person9
10	Ich gehe gerne spazieren	Person10
11	Partner für Nordic Walking gesucht.	Person11

Abbildung 14: Beispiel-Einträge

Es werden sieben verschiedene Arten des Matchings betrachtet, welche in Tabelle 1 für den Eintrag mit der ID 0 aufgelistet sind. Da die *intensive Suche* nicht symmetrisch ist, werden hier unterschiedliche Ergebnisse erwartet. Die Zeit für das Matching wird in Millisekunden gemessen. Hierfür wird folgendes System verwendet: Windows 8.1, i7-2630QM CPU @ 2.00 GHz, 16 GB RAM, Internetverbindung 18/2 Mbit/s.

Jede der folgenden Tabellen enthält die Ergebnisse für den jeweiligen Eintrag. In den Spalten werden die Einträge aufgezählt, welche durch das Matching gefunden wurden. Als Recall wird die Wahrscheinlichkeit, mit der ein relevanter, also ein semantisch ähnlicher, Eintrag gefunden wird, bezeichnet. Precision ist die Wahrscheinlichkeit, mit der ein gefundener Eintrag relevant ist.

Tabelle 1: Ergebnisse für Eintrag ID 0
Ich sehe mir am Sonntag Don Giovanni an. Wer begleitet mich?

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0	3,4,6	3,4	2,5	0.5	0.67
Standard-matching TS	0	3,4,6	3,4	2,5	0.5	0.67
intensive Suche	1936	3,4,6	3,4	2,5	0.5	0.67
intensive Suche TS	2015	3,4,5,6	3,4,5	2	0.75	0.75
sehr intensive Suche	12125	3,4,5,6	3,4,5	2	0.75	0.75
sehr intensive Suche TS	12803	3,4,5,6	3,4,5	2	0.75	0.75
Kategorie-graph	15313	2,3,4	2,3,4	5	0.75	1

Es wird davon ausgegangen, dass für den Eintrag mit der ID 0 die Einträge 2, 3, 4 und 5 relevante Matches sind. „*Konzert*“ und „*Oper*“ werden also als semantisch verwandt angesehen. Der Eintrag 2 mit der gemeinsamen Oberkategorie „*Musik*“ wird nur über den Kategoriegraph gefunden.

Tabelle 2: Ergebnisse für Eintrag ID 1
Ich brauche jemanden der mir beim Einkaufen hilft.

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0			7	0	0
Standard-matching TS	0			7	0	0
intensive Suche	772			7	0	0
intensive Suche TS	1053	6		7	0	0
sehr intensive Suche	12521			7	0	0
sehr intensive Suche TS	12609	6		7	0	0
Kategorie-graph	16709			7	0	0

Da hier nicht zwischen „aktiv“ und „passiv“ oder zwischen „suche“ und „biete“ unterschieden wird, werden Eintrag 1 und 7 aufgrund der Tags „*Shoppen*“ und „*Einkauf*“ als semantisch verwandt eingestuft. Das Match mit Eintrag 6 beim Teilstringvergleich kommt über den gemeinsamen Teilstring „*wirtschaft*“ in den beiden hinzugefügten Tags „*Freizeit-wirtschaft*“ und „*Material- Lagerwirtschaft*“ zustande.

Die semantische Ähnlichkeit mit Eintrag 7 wird von keinem Matching-Ansatz erkannt. Das liegt an den Begriffsklärungsseiten von Wikipedia, wie im folgenden Verlauf noch erklärt wird.

Tabelle 3: Ergebnisse für Eintrag ID 2
Wer möchte mit mir in ein Konzert gehen?

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0			0,3,4,5	0	0
Standard-matching TS	0			0,3,4,5	0	0
intensive Suche	882			0,3,4,5	0	0
intensive Suche TS	917			0,3,4,5	0	0
sehr intensive Suche	12943			0,3,4,5	0	0
sehr intensive Suche TS	12694			0,3,4,5	0	0
Kategorie-graph	16022	0,3,4	0,3,4	5	0.75	1

Die gemeinsame Oberkategorie „Musik“ wird nur bei der Suche im Kategoriegraph gefunden. Da zu „*Musikveranstaltung* kein Wikipedia-Artikel existiert, erfolgt kein Matching mit Eintrag 5. Hierzu wäre ein Teilstringvergleich nötig.

Tabelle 4: Ergebnisse für Eintrag ID 3
Ich gehe gerne in die Oper.

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0	0,4	0,4	2,5	0.5	1
Standard-matching TS	0	0,4,5	0,4,5	2	0.75	1
intensive Suche	1501	0,4	0,4	2,5	0.5	1
intensive Suche TS	1664	0,4,5	0,4,5	2	0.75	1
sehr intensive Suche	14121	0,4	0,4	2,5	0.5	1
sehr intensive Suche TS	14141	0,4,5	0,4,5	2	0.75	1
Kategorie-graph	14996	0,2,4,6,11	0,2,4	5	0.75	0.6

Im Kategoriegraph wurden die Einträge 6 und 11 aufgrund der gemeinsamen Oberkategorie „*Gesellschaft*“ gefunden. Hier reicht die Begrenzung auf die Suchtiefe 4 nicht aus, denn die Oberkategorie ist zu allgemein. Allerdings werden diese im Vergleich zu den anderen Treffern nur mit einem sehr geringen Gewicht angeboten. Die Precision allein ist also nicht unbedingt ein Maß für die Qualität der Ergebnisse.

ID	Eintrag	Autor	Gewicht
4	Ich gehe heute Abend in die Oper und suche noch jemanden der mitkommt.	Person4	2.5
0	Ich sehe mir am Sonntag Don Giovanni an. Wer begleitet mich?	Person0	1.5
2	Wer möchte mit mir in ein Konzert gehen?	Person2	0.125
6	Ich gehe gerne ins Kino und suche eine Begleitung.	Person6	0.0625
11	Partner für Nordic Walking gesucht.	Person11	0.0625

Abbildung 15: Gewicht der Ergebnisse

Tabelle 5: Ergebnisse für Eintrag ID 4
Ich gehe heute Abend in die Oper und suche noch jemanden der mitkommt.

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0	0,3	0,3	2,5	0.5	1
Standard-matching TS	0	0,3,5	0,3,5	2	0.75	1
intensive Suche	1706	0,3	0,3	2,5	0.5	1
intensive Suche TS	2063	0,3,5	0,3,5	2	0.75	1
sehr intensive Suche	13992	0,3	0,3	2,5	0.5	1
sehr intensive Suche TS	15000	0,3,5	0,3,5	2	0.75	1
Kategorie-graph	15698	0,2,3,6,11	0,2,3	5	0.75	0.6

Dieser Eintrag ist dem vorhergehenden sehr ähnlich, daher treten hier wie zu erwarten die selben Ergebnisse auf.

Tabelle 6: Ergebnisse für Eintrag ID 5
Ich interessiere mich für alle Arten von Musikveranstaltungen

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0			0,2,3,4	0	0
Standard-matching TS	0	3,4	3,4	0,2	0.5	1
intensive Suche	857			0,2,3,4	0	0
intensive Suche TS	1028	3,4	3,4	0,2	0.5	1
sehr intensive Suche	12919	0	0	2,3,4	0.25	1
sehr intensive Suche TS	12994	0,3,4	0,3,4	2	0.75	1
Kategorie-graph	17069			0,2,3,4	0	0

Da der Tag „Musikveranstaltungen“ in Wikipedia weder ein Artikel noch eine Kategorie ist, sind hier klar die Verfahren mit Teilstringvergleich im Vorteil. Das Match mit Eintrag 0 bei der sehr intensiven Suche ohne Teilstringvergleich kommt durch die gemeinsame Oberkategorie „*Allgemeine Psychologie*“ von „*Sehen*“ und „*Interesse*“ zustande. Die beiden Einträge sind zwar semantisch verwandt, allerdings sollte das Match aufgrund der beiden Tags „*Oper*“ und „*Musikveranstaltung*“ gebildet werden. Da „*Interesse*“ keine Wikipedia-Kategorie ist, wird dieses Match nicht über den Kategoriegraph erkannt.

Tabelle 7: Ergebnisse für Eintrag ID 6
Ich gehe gerne ins Kino und suche eine Begleitung.

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0				0	0
Standard-matching TS	0				0	0
intensive Suche	1163				0	0
intensive Suche TS	1262				0	0
sehr intensive Suche	12919	10,11			0	0
sehr intensive Suche TS	12471	1,10,11			0	0
Kategorie-graph	16988	3,4,11			0	0

Eintrag 10 hat die gemeinsame Oberkategorie „*Freizeit*“, Eintrag 11 „*Freizeitwirtschaft*“. Das Match mit Eintrag 1 bei der sehr intensiven Suche mit Teilstringvergleich kommt über den gemeinsamen Teilstring in den beiden hinzugefügten Tags „*Freizeitwirtschaft*“ und „*Material- Lagerwirtschaft*“ zustande. Zu Matches mit dem Tag „*Oper*“ kommt es aufgrund der Oberkategorie „*Gesellschaft*“. Alle diese Oberkategorien sind zu allgemein, um von einer semantischen Ähnlichkeit mit dem Tag „*Kino*“ zu sprechen.

Tabelle 8: Ergebnisse für Eintrag ID 7
Ich suche jemanden, der mir mir shoppen geht.

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0			1	0	0
Standard-matching TS	0			1	0	0
intensive Suche	793			1	0	0
intensive Suche TS	852			1	0	0
sehr intensive Suche	12393			1	0	0
sehr intensive Suche TS	12936			1	0	0
Kategorie-graph	18265			1	0	0

Wie in diesem und nächsten Kapitel noch beschrieben wird, werden zum Tag „*Shoppen*“ keine weiteren zusätzlichen Tags erstellt, da die Kategorisierung hier auf eine Begriffsklärungsseite stößt. Dadurch ist kein Matching mit anderen Einträgen möglich.

Tabelle 9: Ergebnisse für Eintrag ID 8
Ich suche eine neue Lauf-Gruppe

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0			9,11	0	0
Standard-matching TS	0			9,11	0	0
intensive Suche	858			9,11	0	0
intensive Suche TS	888			9,11	0	0
sehr intensive Suche	12254			9,11	0	0
sehr intensive Suche TS	12110			9,11	0	0
Kategorie-graph	17507			9,11	0	0

Während der Erstellung der Tags für diesen Eintrag, landen die Wikipedia-Anfragen für „*Lauf*“ und“*Gruppe*“ jeweils auf Begriffsklärungsseiten. Wie bereits erwähnt, ist dadurch kein Match mit Einträgen möglich, welche nicht eines dieser Worte als ursprünglichen Tag besitzen.

Tabelle 10: Ergebnisse für Eintrag ID 9
Wer geht mit mir joggen?

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0			8,11	0	0
Standard-matching TS	0			8,11	0	0
intensive Suche	810	11	11	8	0.5	1
intensive Suche TS	984	11	11	8	0.5	1
sehr intensive Suche	13196	11	11	8	0.5	1
sehr intensive Suche TS	12432	11	11	8	0.5	1
Kategorie-graph	17589			8,11	0	0

Der Eintrag 11 wird über die gemeinsame Oberkategorie „*Trainingsübung*“ gefunden. Da beim Matching über den Kategoriegraph nur Kategorieseiten von Wikipedia berücksichtigt werden und zu „*Joggen*“ nur ein Artikel und keine Kategorie besteht, existiert hier kein Match.

Tabelle 11: Ergebnisse für Eintrag ID 10
Ich gehe gerne spazieren

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0			11	0	0
Standard-matching TS	0			11	0	0
intensive Suche	875			11	0	0
intensive Suche TS	789	6		11	0	0
sehr intensive Suche	13071	6,11	11		1	0.5
sehr intensive Suche TS	12977	6,11	11		1	0.5
Kategorie-graph	17250			11	0	0

Eintrag 11 besitzt die gemeinsame Oberkategorie „*Fußverkehr*“. Eintrag 6 wird fälschlicherweise durch die gemeinsame Oberkategorie „*Freizeit*“ erkannt. Diese werden bei der Suche über den Kategoriegraph nicht erkannt, da der ursprüngliche Tag „*Spaziergang*“ keine Wikipedia-Kategorie ist.

Tabelle 12: Ergebnisse für Eintrag ID 11
Partner für Nordic Walking gesucht.

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0			9,10	0	0
Standard-matching TS	0			9,10	0	0
intensive Suche	916			9,10	0	0
intensive Suche TS	947			9,10	0	0
sehr intensive Suche	12774	6,9,10	9,10		1	1
sehr intensive Suche TS	13001	6,9,10	9,10		1	1
Kategorie-graph	17148	3,4,6		9,10	0	0

Das falsch positive Match mit Eintrag 6 wurde bereits bei den Ergebnissen dieses Eintrags erklärt. Das Matching mit den Einträgen 3 und 4 bei der Suche im Kategoriegraph kommt durch die gemeinsame Oberkategorie „*Gesellschaft*“ zustande. Hier scheint die Suchtiefe abermals nicht genug beschränkt zu sein, da diese Oberkategorie zu allgemein ist.

Die folgende Tabelle 13 zeigt die durchschnittlichen Ergebnisse aller Matchings für alle 11 Einträge. Dabei stehen die Werte für die durchschnittliche Anzahl der Matches und nicht wie zuvor für die ID der Matches.

Tabelle 13: Durchschnitt der Ergebnisse

	Zeit in ms	gefunden	gefunden relevant	nicht gefunden relevant	Recall	Precision
Standard-matching	0	0.64	0.55	1.91	0.14	0.24
Standard-matching TS	0	1.00	0.91	1.55	0.23	0.33
intensive Suche	1105	0.73	0.64	1.82	0.18	0.33
intensive Suche TS	1229	1.36	1.09	1.36	0.30	0.43
sehr intensive Suche	12950	1.27	0.91	1.55	0.32	0.48
sehr intensive Suche TS	13015	1.82	1.27	1.18	0.41	0.48
Kategorie-graph	16673	1.73	1.09	1.36	0.27	0.29

Da durch die meisten der Beispiel-Einträge die Schwachstellen des Systems, beziehungsweise der Wikipedia-Kategorisierung, veranschaulicht werden sollten, fallen die durchschnittlichen Werte für Recall und Precision eher schlecht aus. Auf mögliche Verbesserungen wird im letzten Kapitel eingegangen. Werden alle Einträge welche für keine Art des Matchings einen Treffer erzielen außer Acht gelassen, so ergeben sich folgende Durchschnittswerte für Recall und Precision:

Tabelle 14: Durchschnitt ohne die Einträge 1,6,7 und 8

	Recall	Precision
Standardmatching	0.19	0.33
Standardmatching TS	0.31	0.46
intensive Suche	0.25	0.46
intensive Suche TS	0.41	0.59
sehr intensive Suche	0.44	0.66
sehr intensive Suche TS	0.56	0.66
Kategoriegraph	0.38	0.40

Zu Eintrag 6 gibt es keine relevanten Matchings im Beispieldatensatz. Daher sagen hier Recall und Precision nichts über die Qualität des Matchings aus. Die Kategorisierung der Einträge 1, 7 und 8 trifft bei der Erstellung der Tags auf Begriffsklärungsseiten. Daher kann hier über den Wikipedia-Kategoriegraph kein Matching erfolgen. Eine Lösung hierzu wird im folgenden Kapitel vorgeschlagen.

Insgesamt fällt auf, dass vor allem die *sehr intensive Suche* gute Ergebnisse erzielt. Da allerdings der Zeitaufwand sehr hoch ist, wird im Programm als Standardeinstellung das Standardmatching mit Teilstringvergleich genutzt. Hier sind die Ergebnisse in Anbetracht auf die äußerst geringen Zeitkosten sehr gut. Wird hier kein Match gefunden, wird automatisch die *intensive* beziehungsweise sehr *intensive Suche* angeboten. Durch Behebung des Problems mit den Begriffsklärungsseiten ist sicherlich auch die Suche über den Kategoriegraph eine echt Alternative.

5 Zusammenfassung und Ausblick

Im Allgemeinen kann festgestellt werden, dass die Verwendung von Tags zur Beschreibung der Semantik eines Eintrags in ein Matching-System eine durchaus praktikable Lösung ist. Auch das Kategorie-System von Wikipedia eignet sich sehr gut, um die Semantik von kurzen Texten zu erkennen. Betrachtet man allerdings die Ergebnisse der Evaluation in Kapitel 4, sind Verbesserungen bei der Nutzung von Wikipedia möglich. Im Folgenden sollen einige Ansätze vorgestellt werden, welche eine Verbesserung der Kategorisierung und damit auch ein besseres Matching bewirken könnten.

Ein Teil von Wikipedia, welcher in dieser Arbeit nicht in vollem Umfang genutzt wird, sind die Begriffsklärungsseiten. Diese werden momentan nur dazu verwendet, um das Gewicht eines Tags abzuschwächen. Allerdings besitzen diese Begriffsklärungsseiten einen hohen semantischen Wert. Wird hier die richtige Bedeutung eines Wortes ausgewählt, so grenzt sich der Eintrag von Einträgen, die das selbe Wort mit einer anderen Bedeutung besitzen, ab. Das wäre möglicherweise zu realisieren, indem eine interaktive Auswahl der Wortbedeutung durch den Benutzer erfolgt, sobald eine Begriffsklärungsseite erreicht wird. Im Beispiel aus 4.5 stößt der Tag „*Shoppen*“ von Eintrag 7 auf eine Begriffsklärungsseite (vgl. Abbildung 16). Würde hier die richtige Bedeutung ausgewählt werden, also „*Einkauf(Konsum)*“, wäre ein Matching mit Eintrag 1 möglich. Dieser enthält den Tag „*Einkauf*“

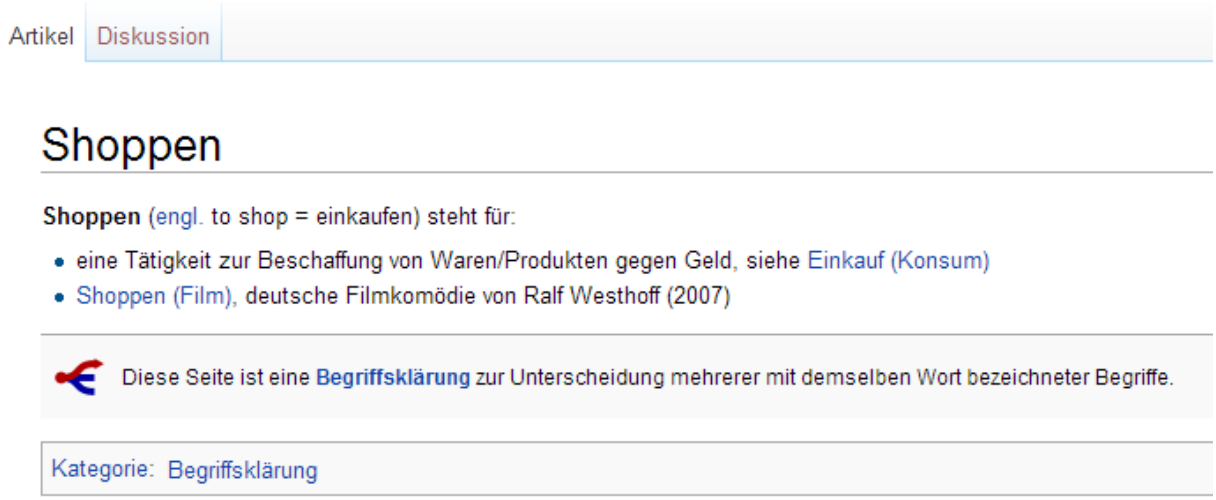

Abbildung 16: Begriffsklärungsseite

Eine weitere Möglichkeit wäre es, dem Benutzer die Möglichkeit zu geben, selbst Tags hinzuzufügen. Dies ist im aktuellen Zustand auch möglich, indem einfach zusätzliche Wörter im Eintragstext eingefügt werden. Allerdings funktioniert das nur, falls es sich dabei um Substantive oder Verben handelt und die Wörter nicht in der Stoppwortliste vorkommen. Eine Alternative wäre, zusätzliche Tags beim Erstellen des Eintrags mit dem Symbol # zu kennzeichnen. Diese würden dann direkt als Tags aufgenommen werden und tauchen nicht im Eintragstext auf. Im Beispiel aus Abschnitt 4.5 könnte Person8 beispielsweise „#Sport“ angeben. So gäbe es ein Match mit dem Eintrag der Person11, da „*Sport*“ eine Oberkategorie von „*Nordic Walking*“ ist.

Ein Ansatz der hier gar nicht betrachtet wird, wäre eine Unterscheidung zwischen Ein-

trägen von Personen, die etwas suchen und welchen, die etwas anbieten. Würde man dies differenzieren, so würden zwei Personen, die nach Hilfe bei etwas suchen, einander nicht vorgeschlagen werden. Suchen allerdings zwei Personen nach einer Begleitung für die selbe Aktivität, wären sie durchaus passende Matches zueinander. Dieses Problem ist also nicht trivial und durch eine einfach Abfrage, ob der Nutzer etwas bietet oder sucht, zu lösen.

Eine Verfeinerung der Matches, welche einfacher zu realisieren ist, wäre die Unterscheidung zwischen aktiven und passiven Freizeitaktivitäten. Ohne diese Unterscheidung bekommt ein Nutzer, der Mitspieler für ein Basketballspiel sucht, Matches von Personen geliefert, die ein Basketballspiel als Zuschauer besuchen. Ebenso werden einem Nutzer, der Personen zum gemeinsamen Musizieren sucht, Einträge von Konzertbesuchern angeboten. Das lässt sich vermeiden, indem der Benutzer angibt, ob seine Aktivität aktiv oder passiv ist.

Zusätzlich wäre es denkbar den Wikimedia-Dienst Wiktionary noch stärker einzubinden. Es könnte beispielsweise nach Übersetzungen gesucht werden. Dadurch wäre ein Match zwischen „*Shopping*“ und „*Einkaufen*“ möglich.

Da das Hauptaugenmerk dieser Arbeit darauf liegt, inwiefern Wikipedia für ein semantisches Matching von Freizeitaktivitäten verwendet werden kann, wurden hier nur die dafür relevanten Aspekte des Matching-Systems implementiert. So wird aktuell der Autor eines Eintrags überhaupt nicht berücksichtigt. Ebenfalls läuft das Programm nur lokal und es ist keine Interaktion zwischen den einzelnen Autoren möglich.

Die Aktualität des Programms ist dadurch gewährleistet, dass der Zugriff auf Wikipedia zur Laufzeit geschieht. Dadurch entsteht kein großer Wartungsaufwand, wie beispielsweise bei Matching-Systemen, welche eine eigene Ontologie oder Taxonomie aufbauen. Die Ontologie wird hier nicht durch die Einträge verändert. So kann diese auch nicht gezielt verfälscht werden. Der Administrator muss also nicht überprüfen, ob falsche Kategorien hinzugefügt oder neue Kategorien falsch eingeordnet werden.

Literatur

[1] J. Grimm. *Berechnung semantischer Ähnlichkeit kleiner Textfragmente mittels Wikipedia*. PhD thesis, Master thesis, Technische Universität Darmstadt, 2009.

[2] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304. Morgan Kaufmann Publishers Inc., 1998.

[3] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2:159–165, 1958.

[4] R. Mihalcea. Using wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL HLT*, volume 2007, pages 196–203, 2007.

[5] D. Milne. Computing semantic relatedness using wikipedia link structure. In *Proceedings of the new zealand computer science research student conference*, 2007.

[6] M. L. Murphy. *Lexical meaning*. Cambridge University Press Cambridge, 2010.

[7] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19:17–30, 1989.

[8] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. 1995.

[9] S. Ross. *A first course in probability*. Macmillan, 1976.

[10] Salton and McGill. Introduction to modern information retrieval. 1983.

[11] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06, pages 1419–1424. AAAI Press, 2006.

[12] M. Ullrich, A. Maier, and J. Angele. Taxonomie, thesaurus, topic map, ontologie-ein vergleich. *Ontoprise GmbH*, 2003.

[13] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[14] T. Zesch and I. Gurevych. Analysis of the wikipedia category graph for nlp applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, pages 1–8, 2007.

Ich erkläre hiermit gemäß § 17 Abs. 2 APO, dass ich die vorstehende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Bamberg, den 16.12.2013

Matthias Düsel